# Mapping the serum proteome to neurological diseases using whole genome sequencing

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-27387-1 OPEN

# Mapping the serum proteome to neurological diseases using whole genome sequencing

Grace Png[1,2✉], Andrei Barysenka[1], Linda Repetto[3], Pau Navarro[4], Xia Shen[3,5,6], Maik Pietzner[7], Eleanor Wheeler[7], Nicholas J. Wareham[7], Claudia Langenberg[7,8], Emmanouil Tsafantakis[9], Maria Karaleftheri[10], George Dedoussis[11], Anders Mälarstig[12,13], James F. Wilson[3,4], Arthur Gilly[1] & Eleftheria Zeggini[1,2✉]

Despite the increasing global burden of neurological disorders, there is a lack of effective diagnostic and therapeutic biomarkers. Proteins are often dysregulated in disease and have a strong genetic component. Here, we carry out a protein quantitative trait locus analysis of 184 neurologically-relevant proteins, using whole genome sequencing data from two isolated population-based cohorts ($N = 2893$). In doing so, we elucidate the genetic landscape of the circulating proteome and its connection to neurological disorders. We detect 214 independently-associated variants for 107 proteins, the majority of which (76%) are cis-acting, including 114 variants that have not been previously identified. Using two-sample Mendelian randomisation, we identify causal associations between serum CD33 and Alzheimer's disease, GPNMB and Parkinson's disease, and MSR1 and schizophrenia, describing their clinical potential and highlighting drug repurposing opportunities.

[1] Institute of Translational Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany. [2] TUM School of Medicine, Technical University of Munich and Klinikum Rechts der Isar, Munich, Germany. [3] Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, UK. [4] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [5] Greater Bay Area Institute of Precision Medicine (Guangzhou), Fudan University, Guangzhou, China. [6] Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden. [7] MRC Epidemiology Unit, University of Cambridge, Cambridge, UK. [8] Computational Medicine, Berlin Institute of Health (BIH), Charité University Medicine, Berlin, Germany. [9] Anogia Medical Centre, Anogia, Greece. [10] Echinos Medical Centre, Echinos, Greece. [11] Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University of Athens, Athens, Greece. [12] Department of Medicine, Karolinska Institute, Solna, Sweden. [13] Emerging Science & Innovation, Pfizer Worldwide Research, Development and Medical, Cambridge, MA, USA. ✉email: grace.png@helmholtz-muenchen.de; eleftheria.zeggini@helmholtz-muenchen.de

Neurological disorders are the leading cause of disability worldwide, accounting for 276 million disability-adjusted life years (DALY) globally in 2016[1]. This burden is continuously increasing with growing and ageing populations[2], emphasising the need for better prevention and treatment strategies. Multiple genetics and genomics efforts have established that these diseases have a substantial genetic component[3,4]. Elucidating their genetic architecture can, therefore, help to forward our understanding of their aetiology by identifying causal disease mechanisms, thus opening a path towards clinical translation.

Due to their heterogeneity and overlapping clinical features, neuropsychiatric disorders such as schizophrenia and bipolar disorder are often misdiagnosed[5], while others with more distinct symptoms, such as Alzheimer's disease (AD), lack effective drugs and accessible biomarkers that can detect early disease[6]. The human serum proteome is an especially valuable resource of potential biomarkers for these highly polygenic disorders. As proteins are often dysregulated in disease, studying protein quantitative trait loci (pQTLs), which are genetic variants associated with protein expression levels, can help to bridge existing knowledge gaps. Most pharmaceutical drugs also target proteins, further increasing their actionability.

By implementing statistical methods that leverage relevant biomedical data, such as causal inference and colocalisation analysis, pQTLs can be used to determine causality and to identify disease pathways. For example, in a study focused on neurologically relevant proteins[7], a pQTL for serum PVR mapping to the *PVR* gene (*cis*-pQTL), was found to be causally associated with AD through Mendelian randomisation analysis. Through similar methods, a recent brain proteome-wide association (PWAS) and pQTL study[8] identified five genes causal for AD at high confidence, of which four were novel. By validating known AD loci and identifying new causal genes, these studies demonstrate proof-of-concept.

Here, we aimed to identify biomarkers of neurological traits and enhance insight into disease pathways, by carrying out a pQTL analysis of 184 neurologically relevant serum proteins. The main advantage of serum proteins is that they are easily accessible, both as drug targets and diagnostic biomarkers. We use whole-genome sequencing (WGS) to capture the entire allele frequency spectrum in 2,893 samples from two Greek population-based cohorts, MANOLIS and Pomak. Association analysis was first carried out individually for each cohort, followed by a meta-analysis. Specifically, proteins were quantified using Olink's proximity extension assay (PEA) and comprised established or potential markers of neurobiological processes. Using WGS, we were able to detect both rare and common pQTL variants. We then investigated the relevance of the discovered pQTLs to neurological diseases and highlight biomarkers of high diagnostic or prognostic potential, identify drug repositioning opportunities, and describe pathways relevant to neurological traits.

## Results

**Protein QTL discovery**. For the 184 neurologically relevant proteins analysed, we detect 214 independently-associated pQTLs ($P < 1.05 \times 10^{-10}$; 'Methods' section) for 107 proteins from the meta-analysis, following conditional testing (Fig. 1 and Supplementary Data 1). Loci were classified into *cis* and *trans*: *cis*-acting pQTLs, which are defined as variants residing within 1 Mb upstream or downstream of the protein-encoding gene, are likely to regulate protein expression directly at the transcriptional level, while *trans*-pQTLs are likely to act through intermediaries to modulate protein levels. We observe 162 (75.7%) *cis*-acting pQTLs for 91 proteins, and 52 (24.3%) *trans*-acting pQTLs for 38 proteins. A total of 22 proteins had both *cis* and *trans*-acting pQTLs (Fig. 2b).

Sixteen proteins have only *trans*-pQTLs, 13 of which have pQTLs only in pleiotropic loci. We find altogether 30 variants arising at known pleiotropic loci, including those near or within *KLKB1, ABO, F12, VTN*, and the HLA region on chromosome 6. These are loci that influence the levels of multiple proteins; the most pleiotropic being loci at *KLKB1* and *ABO*, affecting 11 and 12 proteins, respectively. These have been identified in published pQTL studies and are not restricted to neurologically relevant proteins[9–12]. *ABO* is the most extensively studied among these pleiotropic loci, and is known for its role in blood coagulation processes and determining the ABO blood types. In particular, we detect the missense variant rs8176747 affecting ADAM15, IL3RA, and KIRREL2 protein levels. rs8176747 is among the variants routinely used to determine blood group phenotype[13], which has been associated with multiple diseases, mainly of cardiovascular relevance. As proteins such as ABO are connected to large signalling networks, changes in their structure or expression levels could influence multiple downstream substrates, hence explaining their pleiotropy.

We identify 33 sequence variant-protein level independent associations for 15 proteins that have not been investigated for pQTLs before (Table 1). For the remaining 92 proteins, we identify 72 novel *cis*-pQTL variants, and 15 novel *trans*-pQTL variants, excluding those at known pleiotropic loci. We define novelty if no variants within 2 Mb have been previously reported in serum pQTL studies, or if associations remain significant after conditioning on established pQTLs.

Eight of the proteins we studied here have also been investigated in a pQTL study in cerebrospinal fluid (CSF)[14]. We replicate six of these *cis*-pQTLs in serum: for CD33, GPNMB, LEPR, NAAA, SIGLEC-9, and TDGF1. Additionally, we find novel *cis*-pQTLs for CD33 and GPNMB, and *trans*-pQTLs for NAAA and SIGLEC-9, which had not been detected in CSF. The observed replication of CSF pQTLs indicates that the expression of these proteins in serum and CSF are governed by a shared genetic mechanism.

Of the identified independent pQTLs, 185 (86%) are common-frequency variants (minor allele frequency [MAF] > 5%), 25 (12%) are low-frequency (MAF 1–5%) and four (2%) are rare (MAF < 1%) (Fig. 2a). Eight of the low-frequency or rare pQTLs (all *cis* signals) have not been reported before, despite the proteins having been analysed in past studies, demonstrating the advantage of using whole-genome sequencing-based analysis to capture the full MAF spectrum.

**Gene expression QTL colocalisation**. Colocalisation analysis is used to test if independent association signals from two traits share the same causal variant. When comparing protein with gene expression levels, positive colocalisation is indicative of a shared regulatory mechanism, thereby acting as orthogonal validation. Through testing for colocalisation of neurological pQTLs with gene expression QTLs (eQTLs) from multiple tissues (GTEx), our results also identify disease-relevant tissues where gene expression correlates with serum protein expression. For *cis*-acting pQTLs, analysis was carried out between protein expression and the expression of the encoding gene, in all available tissues. Sixty-four (69%) *cis*-pQTLs colocalised strongly (colocalisation posterior probability 4 [CLPP4] > 0.8; 'Methods' section) with gene expression in at least one tissue, with 11 (12%) in whole blood, and 21 (23%) in various parts of the brain (Supplementary Data 4). This indicates that for these loci, the causal variant influences both gene and protein expression, therefore supporting transcriptional regulation as the mechanism underpinning variation in protein expression levels.

For *trans*-pQTLs, positive colocalisation between a pQTL and an eQTL at a distal gene increases the likelihood that the two gene

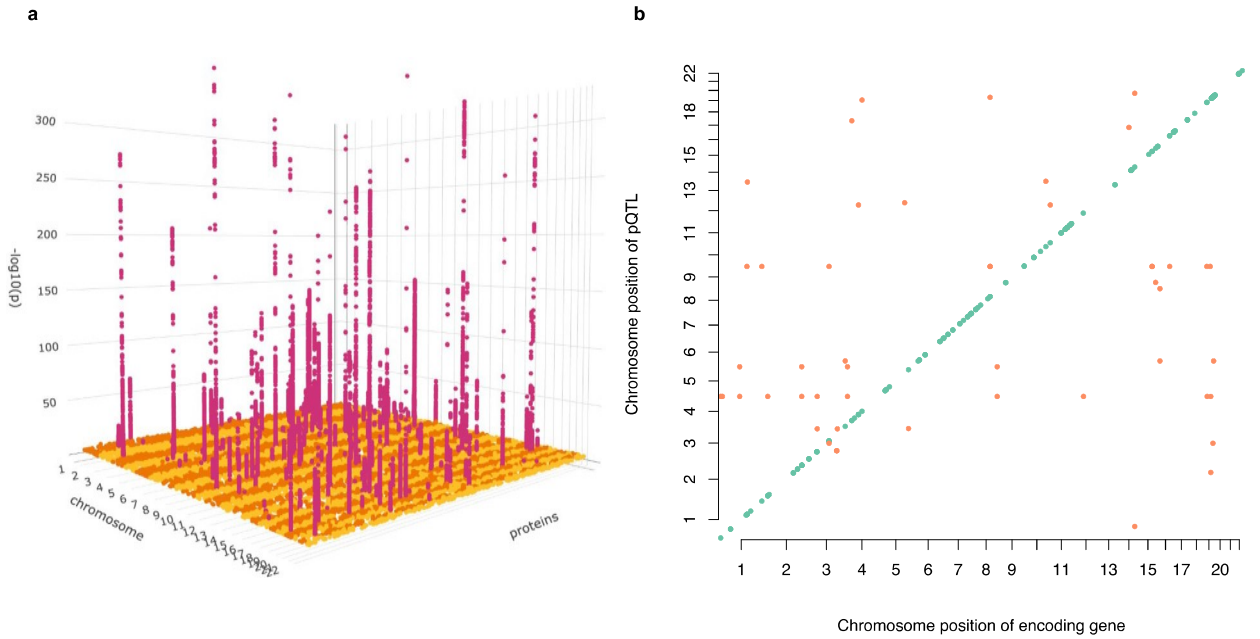


**Fig. 1 pQTL signals for 107 serum proteins from Olink neurology and neuro-exploratory panels. a** 3D Manhattan plot of detected pQTLs. The *x* axis represents each of the 107 proteins; the *y* axis represents the chromosome location of each signal; and the *z* axis represents the −log10 *p*-values of each association signal. **b** Scatterplot of pQTL variant location against the location of the gene encoding the target protein. Each dot represents an independent variant. *Cis*-pQTLs are coloured in teal, while *trans*-pQTLs are in orange.

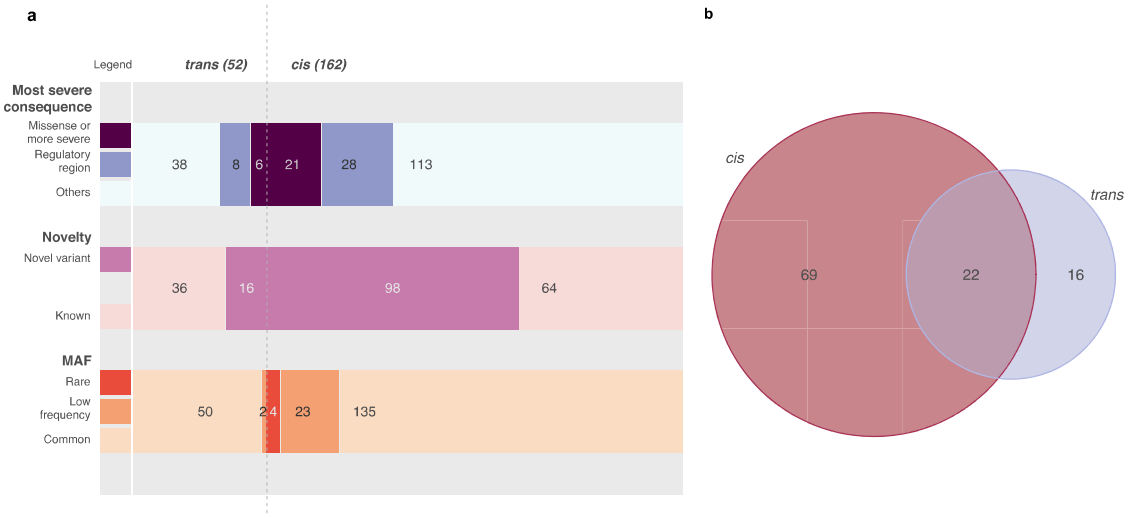


**Fig. 2 Overall genetic architecture of 107 serum proteins of neurological relevance. a** A total of 214 independent variants were detected. *Cis*-acting variants were defined as variants lying within 1 Mb upstream and downstream of the gene encoding the target protein, while *trans*-acting variants are variants that lie outside of this region. Most severe consequence was determined by Ensembl's variant effect predictor (VEP). Effects more than missense included 'stop_gained', 'frameshift_variant', and 'splice_acceptor_variant' in our dataset; 'Regulatory region' variants include '[3/5]_primeUTR_variant', 'TF_binding_site_variant', 'splice_region_variant', and 'regulatory_region_variant'; while'Others' comprises mostly intergenic and intronic variants. Novelty was assessed by cross-referencing published summary statistics from other pQTL studies (Supplementary Data 2). Known pleiotropic loci were not considered novel. Rare, low-frequency and common variants were defined as variants with minor allele frequency (MAF) < 1%, MAF 1–5%, and MAF > 5%, respectively. **b** Number of proteins for which we detected only *cis*-pQTLs, *trans*-pQTLs, or both.

products map to the same regulatory pathway (Supplementary Note 1 and Supplementary Fig. 2). Colocalisation analysis was performed between protein traits and expression of genes within 2 Mb of the *trans*-acting variant. We detect 36 (75%) signals that colocalise with the expression of at least one gene in their vicinity, with three (6%) in whole blood and 30 (62%) in the brain (Supplementary Data 4). As proof-of-concept, we find known receptor-ligand pairs such as a *trans* signal for the KIR2DL3 (killer cell immunoglobulin-like receptor 2DL3) protein colocalising with the expression of *HLA-C* in multiple tissues (22 tissues; CLPP4 > 0.78). KIR2DL3 is an inhibitory receptor for HLA-C, and is responsible for preventing natural killer cells from killing healthy cells[15].

The analysis also enabled the identification of new protein links. For example, we observe a *trans*-pQTL for SMPD1 (sphingomyelin phosphodiesterase; rs10745925; MAF = 0.333; $P = 7.75 \times 10^{-23}$; BETA = −0.2805; SE = 0.0285) that colocalises strongly with the expression of *GNPTAB* in the liver (CLPP4: 0.89), and moderately in

**Table 1 Independent pQTL variants for proteins that are being analysed for the first time.**

| Protein | Variant | MAF | BETA | S.E. | *P*-value | rsID |
|---|---|---|---|---|---|---|
| ADGRB3 | chr6:68956792 | 0.1576 | 0.89 | 0.032 | 8.44E−170 | rs1932618 |
| ADGRB3 | chr6:68962147 | 0.3461 | 0.4947 | 0.0262 | 2.31E−79 | rs3798971 |
| ADGRB3 | chr6:68968025 | 0.3468 | 0.8342 | 0.0225 | 2.83E−301 | rs1953613 |
| CD302 | chr2:159745359 | 0.1016 | −0.4303 | 0.0436 | 5.34E−23 | rs5002908 |
| CD302 | chr2:159773858 | 0.3098 | 0.3731 | 0.0281 | 3.64E−40 | rs1553790820 |
| CDH17 | chr9:133253728 | 0.0918 | −0.6534 | 0.0462 | 1.70E−45 | rs10793962 |
| CDH17 | chr9:133264504 | 0.3431 | −0.3879 | 0.028 | 1.19E−43 | novel |
| CDH17 | chr19:48703205 | 0.4516 | −0.386 | 0.0264 | 2.25E−48 | rs681343 |
| CDH17 | chr8:94194571 | 0.4782 | −0.2672 | 0.0276 | 3.61E−22 | rs56129387 |
| CDH17 | chr8:94130944 | 0.4847 | 0.2889 | 0.0267 | 3.21E−27 | rs1051624 |
| GGT5 | chr22:24232046 | 0.0064 | −2.3071 | 0.1696 | 3.75E−42 | rs200519116 |
| GGT5 | chr22:24235780 | 0.1923 | −0.3614 | 0.0326 | 1.52E−28 | rs6004108 |
| GGT5 | chr22:24247481 | 0.2015 | −0.3049 | 0.0317 | 7.33E−22 | rs5760275 |
| IFI30 | chr19:18172691 | 0.2613 | 0.3604 | 0.0295 | 2.10E−34 | rs273266 |
| IMPA1 | chr8:81652967 | 0.3331 | 0.3338 | 0.0278 | 3.41E−33 | rs2142316 |
| KIR2DL3 | chr19:54744273 | 0.0665 | 0.8024 | 0.0574 | 2.11E−44 | rs10414825 |
| KIR2DL3 | chr19:54743423 | 0.2167 | 0.6973 | 0.0299 | 5.70E−120 | rs11667532 |
| KIR2DL3 | chr6:31272403 | 0.266 | 0.5934 | 0.0307 | 1.71E−83 | rs2524093 |
| KLB | chr17:68883786 | 0.0268 | −0.556 | 0.0849 | 5.79E−11 | rs34931250 |
| KLB | chr4:39431127 | 0.3249 | −0.4173 | 0.0265 | 5.44E−56 | rs2926042 |
| KLB | chr4:39447786 | 0.333 | 0.7642 | 0.025 | 1.17E−205 | rs12513342 |
| LTBP3 | chr11:65572664 | 0.0527 | 0.5989 | 0.058 | 5.49E−25 | rs10896017 |
| LTBP3 | chr11:65575510 | 0.2504 | 0.253 | 0.0299 | 2.68E−17 | rs67924081 |
| NDRG1 | chr5:177412889 | 0.2384 | 0.2707 | 0.0318 | 1.67E−17 | rs2731674 |
| NDRG1 | chr4:186235350 | 0.4738 | 0.2847 | 0.0263 | 2.23E−27 | novel |
| PSG1 | chr19:42929524 | 0.02 | 0.7883 | 0.087 | 1.32E−19 | rs146569565 |
| PSG1 | chr19:42872373 | 0.1525 | −0.3243 | 0.033 | 7.79E−23 | rs60887906 |
| PSG1 | chr19:42881078 | 0.192 | 0.8012 | 0.0267 | 5.72E−198 | rs2005772 |
| RBKS | chr2:27858572 | 0.009 | 1.9199 | 0.1685 | 4.54E−30 | rs140948699 |
| SNCG | chr10:86945549 | 0.2564 | 0.934 | 0.0217 | 3.24E−403 | rs3750822 |
| TPPP3 | chr16:67267204 | 0.0813 | −0.3312 | 0.0483 | 6.86E−12 | rs7200971 |
| VSTM1 | chr19:54062922 | 0.1819 | −0.8967 | 0.0309 | 9.59E−185 | rs8111849 |
| VSTM1 | chr19:54042277 | 0.3968 | 0.8847 | 0.0218 | 4.71E−359 | rs2433724 |

other tissues (CLPP4 = 0.58 [oesophagus mucosa]; 0.57 [stomach]; 0.54 [adrenal gland]). SMPD1 is a lipid hydrolase involved in multiple cell processes; whereas *GNPTAB* encodes subunits of GlcNAc-1-phosphotransferase, which is involved in the synthesis of mannose-6-phosphate (M6P). SMPD1 exists in two forms: secreted and lysosomal. Its lysosomal form is transported via the M6P receptor pathway, therefore supporting the observed SMPD1-GNPTAB interaction. Moreover, we find that the minor allele is associated with a decrease in circulating SMPD1 and an increase in *GNPTAB* expression. This could be a result of increased M6P tagging, which targets a disproportionate amount of the enzyme to the lysosome rather than the secretory pathway. Secreted and lysosomal SMPD1 are likely to play distinct roles in the body[16], and abnormal levels of the secreted form have been implicated in age-related neurodegenerative conditions[17] including Alzheimer's disease[18] and amyotrophic lateral sclerosis (ALS)[19]. We, therefore, identify a locus at *GNPTAB* that coregulates secreted SMPD1 levels and *GNPTAB* expression, pinpointing a possible mechanism behind SMPD1-related neuropathological disorders.

**Heritability**. To estimate the narrow-sense heritability of the protein traits studied, the proportion of variance explained (PVE) by all variants across the genome was calculated using GCTA GREML[20] for each protein. Using a single-component approach, WGS variants explained a median of 33.3% of variance in serum protein levels, with the highest observed heritability observed for CD33 ($h^2 = 87.2\%$). Another three proteins had high heritability of more than 80%: TDGF1 (85.4%), VSTM1 (82.8%), and LAIR2 (82.3%). Conversely, some proteins had very low heritability estimates of $h^2 < 5\%$: IKZF2 (4.9%), RNF31 (4.4%), and EPHA10 (0.001%).

We observe that for all four proteins with $h^2 > 80\%$, the pQTLs colocalised with gene expression QTLs in multiple tissues, indicating regulation at the transcriptional level; therefore, the high observed $h^2$ values are likely to mirror genuine high heritability. There are, however, other non-mutually exclusive reasons that can drive very high or low estimates: (1) Variants that alter the binding specificity of the Olink antibody but not the quantity of protein may produce inaccurate heritability estimates; and (2) Known and unknown biases of single-component GREML approach, which tends to overestimate $h^2$ when causal variants are common, and underestimate $h^2$ when causal variants are rare[21] (Supplementary Fig. 1).

**Link to disease outcomes**. To explore the biological relevance of the pQTLs, we carried out colocalisation analysis with neuropsychiatric traits using data published by the Psychiatric Genomics Consortium (PGC), as well as other neurodegenerative traits, using publicly available summary statistics from recent large GWAS meta-analyses (Supplementary Data 5b). We also studied colocalisation with signals for pain-related traits that have been proven to have a neuropathic component, such as chronic back pain[22] and osteoarthritis[23]. A total of 15 protein–trait pairs colocalised with human disease signals, suggesting a role for the protein in mediating disease. These results are summarised in Supplementary Data 5a.

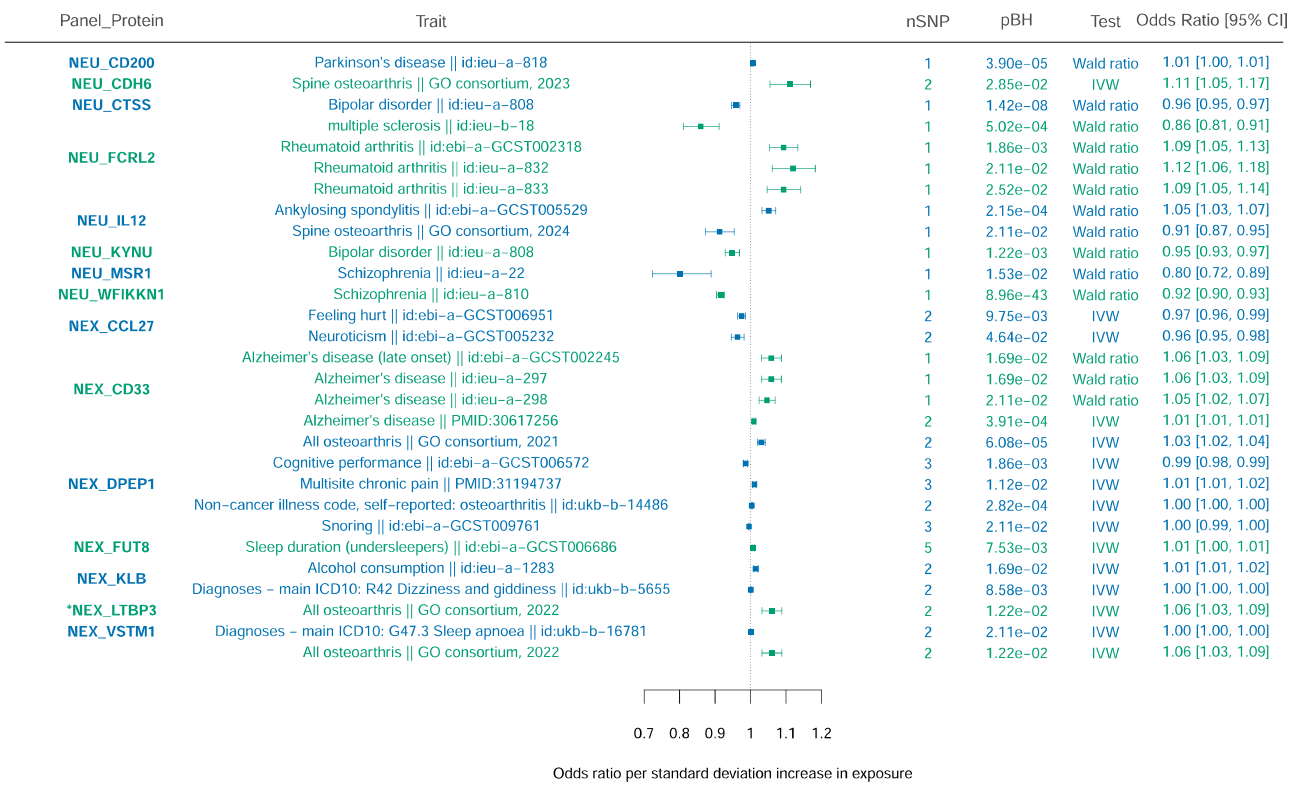


**Fig. 3 Causal protein-disease associations identified using two-sample Mendelian randomisation.** We investigated the causal effect of serum proteins (exposure) on various neurological traits (outcome), indicated in the first two columns in the plot. PubMed IDs (PMIDs) are given where manually downloaded summary statistics were used; other IDs are those as given in MRBase (https://gwas.mrcieu.ac.uk/). The number of variants used in the analysis are given in the 'nSNP' column. The 'pBH' column contains the FDR-adjusted (Benjamini–Hochberg) *P*-value for each test. Protein-trait pairs with only one variant were analysed using the Wald ratio method, while those with more than one variant were analysed using the inverse variance-weighted (IVW) method. Data are represented as mean odds ratio ± SEM. *Additional signal arising from analysis using only *cis*-pQTLs as instrumental variables.

We applied two-sample Mendelian randomisation (MR) for the 107 proteins for which we detect pQTLs, and 206 neurologically relevant and behavioural traits. In contrast to colocalisation, the objective of MR is to look for causal effects of proteins on neurological phenotypes. Using both *cis* and *trans*-acting pQTLs, fifteen proteins were found to be causal for at least one trait, and we detect significant causal effects for 25 unique protein–trait pairs (Fig. 3 and Supplementary Data 6a).

We replicate multiple known associations between protein and disease from the colocalisation and MR analyses. These include LEPR (leptin receptor) and migraine[24], LTBP3 (latent-transforming growth factor beta-binding protein 3) and osteoarthritis[25], FLRT2 (leucine-rich repeat transmembrane protein) with bipolar disorder[26], and PLXNB1[27] (plexin-B1) and PLA2G10[28] (group 10 secretory phospholipase A2) with schizophrenia.

The analysis also identified new protein-disease relationships. Notably, the strongest causal association was found between serum WFIKKN1 and schizophrenia ($P_{adj} = 9.12 \times 10^{-43}$); WFIKKN1 (WAP, Kazal, immunoglobulin, Kunitz and NTR domain-containing protein 1) has not been associated with any neuropsychiatric disorder to date, but is highly expressed in the brain (GTEx) and regulates the activity of several growth and differentiation factors[29]. Similarly, we find new evidence that serum VSTM1 is causally associated with sleep apnoea ($P_{adj} = 2.03 \times 10^{-2}$). VSTM1 (V-set and transmembrane domain-containing protein 1) is a cytokine that promotes the differentiation of helper T-cells (TH17), which are often implicated in autoimmune disorders that may develop secondary to sleep apnoea[30,31].

The overarching aim of this study was to identify protein biomarkers that may be used in the prognosis, diagnosis, or treatment of neurological diseases. Here, we highlight various potential disease markers that are supported by multiple lines of evidence.

**GPNMB as a biomarker for Parkinson's disease**. We identified a *cis*-pQTL that is associated with decreased levels of serum GPNMB (transmembrane glycoprotein NMB; rs7797870; MAF = 0.4286; $P = 7.01 \times 10^{-50}$; BETA = −0.2109; SE = 0.0247) and colocalises with a known Parkinson's disease (PD) locus[32] (CLPP4 = 0.86) (Fig. 4b). *GPNMB* has been highlighted as a susceptibility gene in large PD meta-analyses[32] and has been proven to be upregulated in the brains of PD patients and in mice with induced lysosomal dysfunction[33]. In addition to its connection to PD, we present new evidence showing that serum GPNMB shares a causal variant with *GPNMB* gene expression in both whole blood (CLPP4 = 0.79) and brain tissue (basal ganglia CLPP4 = 0.70; cortex CLPP4 = 0.74; anterior cingulate cortex CLPP4 = 0.83). This not only implies that GPNMB expression is regulated transcriptionally by the pQTL, but also that its expression in the blood and brain are mediated via a shared mechanism. This is supported by previous research showing that tissue GPNMB is able to shed its ectodomain and enter circulation[34]. The lead variant rs75801644 explained 7% of variance in antibody binding for serum GPNMB. Importantly, the identification of serum GPNMB levels as a potential marker of PD is significant as current diagnostic biomarkers are mostly found in the CSF. As serum biomarkers are much less invasive to